DOCUMENT RESUME

ED 154 672 HE 009 943

AUTHOR	Kenoyer, Charles E.
TITLE	Planned Variations Study. Volume III: Design and Analysis Considerations for a Longitudinal Field Experiment.
INSTITUTION	System Development Corp., Santa Monica, Calif. Studies and Evaluation Dept.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
REPORT NO	TM-5857/003/00
PUB DATE	31 Jan 78
CONTRACT	300-76-0539
NOTE	53p.; For related documents see HE 009 941-946
EDRS PRICE	MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS	*Compensatory Education; *Disadvantaged Youth; Evaluation Criteria; Evaluation Needs; *Intervention; *Longitudinal Studies; Models; *Program Evaluation; Research Projects

ABSTRACT

System Development Corporation conducted an extensive project for compensatory educational intervention for older disadvantaged youth in secondary and postsecondary education. Two intervention strategies were developed, and this report discusses the design of a longitudinal study to evaluate the effects of models growing out of those interventions, along with analytic procedures to answer the issues addressed in such a longitudinal study. (Author/MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

System Development Corporation

ED154672

PLANNED VARIATIONS STUDY

DESIGN AND ANALYSIS CONSIDERATIONS FOR A LONGITUDINAL FIELD EXPERIMENT

VOLUME III

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATIO ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

HEW 9 943

JANUARY 31, 1978

TM-5857/003/00

PLANNED VARIATIONS STUDY

VOLUME III: DESIGN AND ANALYSIS CONSIDERATIONS FOR A LONGITUDINAL FIELD EXPERIMENT

Charles E. Kenoyer

Studies and Evaluation Department
System Development Corporation

January 31, 1978

TM-5857/000/03

The work on which this publication is based was performed pursuant to Contract 300-76-0539, as amended, with the U.S. Office of Education, Department of Health, Education, and Welfare. Views or conclusions contained in this study should not be interpreted as representing the official opinion or policy of the Department of Health, Education, and Welfare.

CONTENTS

	Page
INTRODUCTION	1
SPECIFICATION OF DESIGN AND ANALYSES	5
Identification of Issues	5
Specification of Analytical Questions	6
Examination of Alternative Approaches	7
Selection of Design and Analysis Approaches	9
DESIGN CHARACTERISTICS	11
Cross-Sectional Aspects	13
Blocking	13
Treatment/Control Assignment	16
Comparison Sites	17
Longitudinal Aspects	18
ANALYSES OF STUDENT DATA	21
Definitions	21
Units of Analysis	23
Indicators of Program Success	24
Adjustments or Corrections to Scores	26
Uncorrected Post-Treatment Scores	26
Indices of Growth Using Pre-Treatment Measures	27
Manipulation Versus Observation	29
OTHER CATEGORIES OF EVALUATION DATA	32
Process	33
Assessment of Implementation	33
Context	34
Spinoffs/Spillover	35
SUMMARY AND CONCLUSION	36
APPENDIX: Study Questions	A-1

INTRODUCTION

During the period of September 1, 1976, through January 30, 1977, System Development Corporation (SDC) conducted an extensive planning effort under a contract with the U. S. Office of Education (USOE). Called the Planned Variations Study, a major dimension of the project involved the conceptualization and planning of models for compensatory educational intervention at the secondary and postsecondary levels. The project also included the planning of a longitudinal evaluation of the models, incorporating the design of an evaluation and analytic procedures to be used, the methodology for executing an evaluation study in a variety of field settings, and the selection of appropriate instrumentation to determine the effects of models on participating students.

The genesis for the Planned Variations Study arose from earlier evaluations of two programs operated by the Bureau of Higher and Continuing Education (BHCE), the Upward Bound program for secondary students and the Special Services program for postsecondary students. Based on the results of these earlier evaluations, which had indicated that both Upward Bound and Special Services were achieving certain of their objectives but could benefit from the development of alternative models, the Office of Planning, Budgeting and Evaluation (OPBE) contracted with SDC to carry out the planning activities outlined above.

In the 17 months of the Planned Variation Study, SDC performed the following tasks, each of which is described in a separate report:

- Developed two intervention strategies for compensatory education projects to meet the unique needs of disadvantaged secondary and postsecondary students.
- Developed the design of a longitudinal study to evaluate the effects of models growing out of the two intervention strategies, along with analytic procedures to answer the issues to be addressed in such a study.
- Developed a field implementation plan for a longitudinal study, by which intervention models could be implemented in a variety of settings and data collected as called for in a longitudinal study design.

- Developed a set of recommendations for instruments to measure student outcomes in a longitudinal evaluation, based on an analysis of available measures in the cognitive and affective domains.
- Developed an index of components of existing projects and a system for regularly adding to and maintaining the index.

The purpose of this paper is to describe problems that arise in a certain category of research project, and to present some thoughts on ways of dealing with such problems. The kinds of design and analysis problems that merit inclusion in a paper of this kind are rather specific--general problems of design and analysis have been dealt with effectively in standard texts* on the subject, and so need not be discussed here. The emphasis, then, is on specific steps in the design and analysis tasks as they fit into the kind of research project exemplified by the Planned Variations Study (PVS).

In order to be more explicit about the scope of the discussion it will be helpful to describe the defining characteristics of PVS--i.e., those that impose constraints on its design and analysis tasks. First, the philosophy underlying the authorization of the study called for examination and comparison of educational interventions that would be systematically varied, rather than observed as they "naturally" vary in the field. This field experimental aspect of the study is its most influential characteristic in determining the kind of design required. Many of its implications fit into the "general design principles" rubric referred to previously, but political-technical tradeoffs did arise in PVS regarding the kinds of comparison sites to be used, and so this design problem is discussed in the appropriate section.

A second characteristic of PVS is its period of implementation. It was specified in the RFP that the implementation would be for three years at most. This limit requires that outcome measures to be used during the study be of an "intermediate" kind, rather than include such ultimate indicators of success as income, profession, etc. that could be examined after a longer period. The three-year period

*For example, B. J. Winer, Statistical Principles in Experimental Design, McGraw-Hill, 1962; M. R. D'Amato, Experimental Psychology, McGraw-Hill, 1970.

also provided opportunities that would not be present in a one-year study. Even after consuming one of those years in a formative activity, two years of program performance could be appraised. This kind of appraisal is clearly needed in such studies in order to determine whether the first year is typical, whether "seasoning" leads to more effective operation, or whether performance of the program drops off as its novelty fades.

Third, specific objectives of PVS, stated in the RFP, called for certain comparisons to be made. These had to be worked into the design, but were pretty well covered by general design considerations and so had little impact on the resulting design.

Fourth, the study was multi-faceted, and consisted of parts that had to be done sequentially. Summative evaluation activities--i.e., summative data collection and analysis--would begin fairly late in the sequence, after the projects had been fully implemented. This implementation would have been preceded by a formative evaluation, which would not have occurred until development of the program packages was complete. Planning of the design and analysis could, of course, begin at the outset, but some of the details of these plans could not be filled in until the shape of the programs began to crystallize.

Preceding even program development was an activity (called for specifically in the RFP) that would determine the parameters of the population to which PVS should address itself. The influence of information from such an activity is all-pervasive in the sequence of activities that constitute the project. Decisions about the structure of the programs, the kinds of sites to be chosen for implementation, the variables that define the design, and the outcome variables that serve as indicators of success are all included. Suggestions about the contracting of this kind of activity are offered.

Finally, an important aspect of the study is its contractual character. The design of a study can be accomplished in a number of ways, including a sudden flash of insight followed immediately by a trip to the laboratory to conduct a pilot study. When a contract is involved, however, it places additional

requirements on the researcher. The specifics of the additional requirements depend upon the particular kind of contract under which the study is done, but the usual arrangement is a cost-plus contract, and so that will be assumed here. Cost-plus contracts inherently place the greater risk on the customer, rather than on the contractor, and so typically require more intermediate indicators of progress than do fixed-price contracts. For this reason, the development of design and analysis plans needs, at some cost in efficiency, to include check points along the way. There probably are no general rules for the spacing of these checkpoints, but some considerations are cited in the next section of this paper.

SPECIFICATION OF DESIGN AND ANALYSES

This section describes a process by which a study design and an analysis plan can be generated. The process operates "from the top down," beginning with the broad issues to be addressed by the study, proceeding to more specifically defined questions, identifying and examining alternative approaches to answering the questions, and, finally, selecting the design and analytical approaches that will be used. The steps provide a number of checkpoints along the development path, where the customer can review what has been accomplished and provide direction at each point. (It is important to recognize that the customer may disagree with the entire approach to developing the design and analysis plan. In such cases, it is not clear what alternative procedure should be followed, or what authorizes or "drives" the analyses, but the importance of obtaining agreement to the process should be emphasized.)

IDENTIFICATION OF ISSUES

Issues to be addressed by a contractual study originate primarily in two documents: the Request For Proposal (RFP) and the proposal itself. The issues gleaned from these sources provide, at least, a good starting point for the development of a comprehensive list of issues. Such a list should be augmented, of course, by any additional issues that have crystallized between the writing of the proposal and the creation of the list. The major objective of this step is comprehensive coverage; no issues should be discarded for any reason except that they lie outside the intended scope of the study.

After the list is completed, confirmation needs to be obtained. The kind of review of the list that is appropriate for a given study depends upon the level and kinds of resources available. A useful example is the review that was conducted in the Planned Variations Study (PVS): the Policy Research and Advisory Group (PRAG) for PVS was asked to review the list and to suggest additions, modifications, or deletions (if an issue seemed to be out of scope). The inputs obtained from the PRAG were used to revise and refine the questions that expressed the study issues.

It should be noted that the procedure was successful in obtaining inputs, but there was some difficulty in that some of the respondents expressed discomfort in responding to these general questions. They would have preferred to react to the entire research design. This objection does not invalidate the procedure, however; it only indicates that reviewers are uncomfortable when asked to pass judgment on an intermediate product where the implications for how the research is to be conducted are not clear. There are at least two ways to deal with the problem. The reviewers might be reassured by seeing the entire review process charted beforehand, so that the opportunities for further review would be clear. It would also be reasonable to defer review until the next step in the development of design and analysis. The latter approach saves some review time and avoids the dissatisfaction that can result for the reviewers, but it sacrifices input from the reviewers.

SPECIFICATION OF ANALYTICAL QUESTIONS

The second step elaborates the broad issues, deriving a set of questions that suggest approaches to their answers. In general, the second level of question moves closer to "operational" language; i.e., it is stated in terms of differences between scores on one or more variables. Derivation of the analytical questions can be trivial in some cases and involve a great deal of interpretation and elaboration in others, depending upon the complexity of the broad issue, the clarity with which it is stated, and the difficulty of translating all of its terms to observable variables. While the first-level set of questions should reflect as nearly as possible the language in which the study was authorized (so that their origins will be clear) the second-level questions should communicate as clearly as possible the variables and conditions being manipulated, so as to facilitate the next steps.

The task is one of literary interpretation and no algorithm can be specified for deriving the second set of questions from the first. It is reasonable to specify that for every question at the first level there will be at least one at the second level, but the links from the first-level questions to the second-level ones cannot be specified for the general case. Perhaps it is useful to refer to an example, shown in the Appendix, which lists the set of questions derived for PVS.

The broad question (1.1) is entitled: EFFECTS OF TREATMENTS RELATIVE TO CONTROLS. One subdivision of this question differentiates Same-School Controls (1.1.1) from Other-School Controls (1.1.2). This distinction did not appear in earlier versions of the list because it did not arise as a concern until design issues were discussed. Within each of these subdivisions, a second breakdown of the question appears, based upon the kind of outcome measure to be used: (1.1.1.1) Mean Outcome Scores, (1.1.1.2) Adjusted Outcome Scores, and (1.1.1.3) Success Rates. Under other questions, still other distinctions become important.

This hierarchically structured set of questions serves several purposes. First, it serves as a geneology for each analysis to be performed, facilitating the later interpretation of results in terms of the initial study issues. Second, it lays out the total set of issues and analyses so that priorities can be established, thus facilitating orderly elimination of analyses if a cutback is required. Third, it provides a visualization of the logical interrelationships among the analyses and design characteristics. Fourth, it makes "gaps" in the structure apparent, and directs attention to additional questions needed to complete the logical structure. Finally, it serves as the logical link between broad study issues and the design and analyses that are subsequently developed.

EXAMINATION OF ALTERNATIVE APPROACHES

For each of the specific questions one or more analyses must be planned. At this stage (and, to some degree, in the previous one) it is possible to evaluate the feasibility of an analytical approach. If no designs/analyses can be worked out that will yield answers to one or more questions, these questions must be rephrased or abandoned. As was indicated above, it is also possible to add to the question list on the basis of design or analysis considerations.

This task involves establishing the strategy for the study. Some of the variables are amenable to manipulation, but only a limited number of conditions can be specified in the study design. As the limitations on the design are clarified, variables can be prioritized, and the most important can be selected for manipulation. Variables that seem to be important contributors to study outcomes may also be allowed to vary naturally and may be handled analytically, as in a regression analysis.

For PVS, the design was initially described in the proposal as a blocks x treatment design at each level--secondary and postsecondary--in which the treatments were unspecified program variations and the blocks were unspecified characteristics of sites. The first step in refining the design was to identify the variables that were to define the blocks and treatments.

The initial design provided a starting point for identifying blocking variables. Even though the four blocks specified in the proposal were not taken to be fixed, it was assumed that approximately that number would be used. After the potential blocking variables had been called and prioritized, the number of blocks selected in the first update of the design was four. A separate design resulted for each level, since the variables selected were not the same for secondary and postsecondary levels. The selected variables were documented and discussed with the customer for confirmation.

The treatments were to be defined as types of program interventions, with two intervention models defined at each level. Like the blocking variables, the treatments were left undefined at the outset, and were to be specified as program development progressed. Although development of new programs was a central part of the proposed study, it affects design only at the level of specifying what the treatments are, and not structurally. Consequently, program development will not be discussed here.

The RFP emphasized that effectiveness of programs should be qualified in terms of the type of student to be served. The identification of viable student variables was a major part of this problem.

- Could a given variable be measured?
- Would measuring it violate any individual rights?
- If we had the evidence, would it later be useable for selection?
- Was there evidence that the variable influenced the outcome variables on which the interventions were to be evaluated?

But even having identified a half-dozen or so of these variables, there was still the important task of deciding how their effects could be assessed.

If the variables were to be manipulated, their levels would become part of the study design. Alternatives include the following:

- Blocking sites on student characteristics
- Blocking on students within sites

Blocking sites would either have to be done with each blocking variable factorially orthogonal to the rest, with an incomplete factorial design (i.e., with some combinations missing), or with some or all of the variables confounded. Each of these alternatives was ruled out for one reason or another, which will be described more specifically in a later section. The remaining alternative was to deal with the variables in the analytic techniques rather than in the design.

In this way each variable (or category of variables) and each issue is examined for feasibility, and a method of dealing with it is identified. If no method can be identified, the variable must be discarded or the issue must be revised or discarded. After a feasibility check there are still potentially a variety of viable alternatives. The final step is the selection of the one to be implemented.

SELECTION OF DESIGN AND ANALYSIS APPROACHES

One logical method for selecting the design and analysis approach to be used from the set of alternatives generated in the preceding step is to establish criteria for evaluating them and apply those criteria in an orderly way to generate a decision. In the real world, however, the customer's criteria may not all be known to the researcher, and final decision power usually lies with the customer. Alternatives therefore need to be presented to the customer along with any arguments and rationale that can be brought to bear on the decision problem.

The process referred to here does not, in fact, consist of four steps to be carried out in a fixed order. At any of the latter steps it may be necessary to

revise an earlier step; examples of this process were given above. This retrogressive sequence of events nonetheless trends in the direction described here: the issues are initially defined, they are analyzed into component questions, approaches to answering the questions are generated, and the preferred approach is selected for use.

DESIGN CHARACTERISTICS

The purpose of this section is to deal with specific design characteristics for field experiments, rather than general design principles of the sort that are described in statistical textbooks. Consequently, the issues to be discussed must be related more directly to the characteristics of PVS than those of the preceding section were. Perhaps the best definition of the essential characteristics of PVS is embodied in the statement of the objectives of the study given in the RFP. These are quoted below:

Objectives of the Study

The major objectives for all three phases of the study can be specifically summarized as follows:

- Identify appropriate target populations for the planned variations design.
- Diagnose the likely educational and educationally related needs of those populations focused on increasing motivation and educational attainment for entrance to and success in postsecondary education.
- Search the available research on educational skills development, motivation, and theories of educational attainment. From the search, develop models, or "packages," of educational attainment which specify the longitudinal, causal relationships among different input, process, and output variables so that variables can be manipulated to improve program impact.
- Identify for possible use any promising existing models of compensatory educational intervention that may be used or adapted for the planned variations design.
- Plan the implementation and design the evaluation of the models, or planned variations, in a randomized, controlled field test.

The evaluation design should include the following objectives:

- Compare the models' effects with each other, with the effects of regular schooling, and with the effects of the existing program.

- Evaluate the cumulative effects of different treatments, under different exposure durations.
- Determine the differential effectiveness of the models after one, two, or at most three years of implementation.
- Identify and document successful models, and compare them with unsuccessful models to determine model components that differentiate success from failure. Analyze the cost-effectiveness of the models.
- Determine the longitudinal impact of the planned variations on verbal and mathematics achievement, study skills, and motivation for academic success and further education.
- Study the relationship among regular school or college expenditures, supplemental compensatory expenditures, and program impact with particular emphasis on determining the minimum supplemental expenditures necessary to insure some measure of compensatory program success.
- Assess the long-term effects, or "fade-out" of effects, after treatment termination and those experiences, or lack thereof, associated with each phenomenon.
- Compare and contrast the "value added" achievement production functions for the several models to be tested; compare with the control group, and compare with the existing program.

The first four of the above-stated objectives deal primarily with other parts of the PVS project, but some have implications for design. The fifth objective is the charter for the evaluation effort; it is elaborated in the eight evaluation objectives that follow it.

The objectives deal with a number of facets of the study design, including which models (or programs) are to be compared with each other, schedule of the evaluation, cost-effectiveness, (including "value added" analyses), expenditure analyses, cumulativeness of program effects, and fade-out or retention of effects. The interest in the target population expressed in the first objective was later elaborated into a design-relevant issue. The general question that represented

the overall purpose of the study was: "What programs are effective for what kinds of students, in what kinds of contexts?" Some restrictions are necessary at the outset, to clarify the question. First, "kinds of students" can only vary within the range of the disadvantaged category. The customer was not interested in either the existing non-disadvantaged student or some hypothetically-possible but nonexistent type of disadvantaged student. Similarly, the customer was interested only in existing rather than hypothetical types of institutions (contexts), and preferred to deal with institutions like those currently served by Special Programs. The programs to be examined, however, were not be be selected from existing ones, but were to be developed. The answer(s) to the question could not, therefore, be generalized to any population of programs (of which the PVS sample would be representative), but would be specific to those generated and tested. (The development of programs was not a part of study design, and so is not treated here.)

CROSS-SECTIONAL ASPECTS

As the requirements of the design crystallized, it was clear that both context and student characteristics would have to be considered in laying out the design, as would the kinds of groupings needed to obtain the requisite comparisons. For convenience, blocking and treatment conditions are discussed separately.

BLOCKING

The customer was aware when the RFP was released that, although student characteristics were likely to be an important factor in program outcomes, the specific characteristics that should be taken into account were not known. It was therefore specified in the RFP that a Target Population Analysis (TPA) would be performed to determine the composition of the target group. The steps for blocking on student characteristics, then, were:

- Identify the student characteristics that define who should be served by the Special Programs and determine which of these are related to program success.
- Decide which of these can feasibly be included in the study design, and how.

It may be instructive to examine the first step more closely. The key phrase "who should be served" (emphasis added) extends the task beyond the bounds of scientific data gathering, analysis, and reporting.

It is important to recognize that the role of the researcher does not extend to policy making; it is limited to data-based recommendations as to the most effective implementation of existing policy.

After a great deal of discussion involving both the contractor and the customer, it was decided that the existing legislature and the regulations that served to interpret them were to constitute the basis for the definition of the target population, with some of the crucial terms of those documents being clarified through analysis of the appropriate data base.

A range of variables, such as income level, adequacy of high-school preparation for postsecondary education, and handicapped status were abstracted from the legislation and regulations.

Having established this context, the task was to locate existing (published) data relating target-student characteristics to measurable outcomes. What was needed was a data base with (1) a nationally representative sample, (2) longitudinal data, and (3) the mandated characteristics of the target population. It was obvious from the outset that this ideal data base could not exist unless it had been in reports submitted to major educational bureaus or offices within the Government, in which case it would already have been known to the customer. Nonetheless, an attempt was made to piece together data relevant to the target population. Even though it would not be possible to correlate a given portion of such data to other parts, the outcomes were rough indices such as high-school completion rates, and there were relatively few student characteristics that both could be used as predictor variables to those outcomes and fit the legislative constraints on the target population. At this point it was necessary to decide whether to terminate the search and use the data with all of its shortcomings or to extend the scope of the search. It was the customer's judgment that a further search would be necessary.

The scope of this search was extended to include unpublished data bases that were available on tape. It was felt that the data base from the National Longitudinal Study (NLS) of the Class of 1972, properly subset, would serve as a surrogate for the data base described previously. The customer decided at that time to perform a standard set of analyses (mostly cross-tabulations) on the data, rather than amend the contract to oover the new work.

The point of this rather long historical digression is that the specifics of design rather than the general principles constitute the major difficulties. At the outset--i.e., at the time of the proposal--it was known that the design would have to be in the general form of a blocks x treatment factorial design. There was even a rough idea, subject to some modification, of the allowable number of factors. What was not known was what the blocking factors and treatment factors would be. Nearly a year had passed when the blocking, treatment, and control factors were all identified and agreed to by the customer. The bases for the selection of these variables was partly political and, in most instances, involved the mental weighing of tradeoffs by the customer. It is not clear how such frustrating delays can be entirely avoided, but perhaps the following discussion will be of some help.

In any study that involves experimental manipulation or experimental blocks, it is important to know before the design is laid out what variables are important for such a study. This objective would best be served by a separate exploratory effort in which data are gathered specifically for the purpose and in which analytical resources (computer time, etc.) are allocated to the task. Such a study would have as its objective the identification of measurable variables that are potential predictors of success (after first establishing an acceptable definition of success) in the absence of special interventions, and to assess the strength of the predictors. If practically useful predictors were identified, they could be used as avariables in the design of a subsequent study. They could also serve as guidance in the reformulation of legislation and regulations, if desired.

The point to be made is one of emphasis. It is a simple matter to lay out a design diagram in which treatment x blocks combinations define the cells, on the

basis of a general knowledge of design principles. Fleshing out the parameters of the design, however, requires a sophisticated combination of empirical data and knowledge of the policy issues. In the final analysis, the decision needs to be made by the customer after the empirical reality has been assessed. In all likelihood, the necessary data for a thorough analysis of the target population still will not have been collected when another field experiment is initiated, unless a target population study is deliberately designed for that specific purpose.

The rest of the blocking problem consists of prioritizing potential blocking variables and selecting as many of them as are cost-feasible to specify the design. The decision requires a balancing of cost, potency of variable, and policy importance, and must ultimately be made by the customer. The list of variables generated in PVS was created in an orderly way (documented elsewhere), but without much basis in empirical data. Subsequently it was decided by USOE that the design should include, at the secondary level, a condition in which students would be segregated by level of need, and the customer was to derive the predictive system from NLS data that would be used in creating a neediness scale. This analysis, or an elaboration of it (perhaps on new data) would be an appropriate part of the approach described just above.

TREATMENT/CONTROL ASSIGNMENT

A critical attribute of PVS was its field-experimental character. Inclusion of control groups was planned from the start, beginning with the proposal, but some doubt emerged as to the possibility of establishing control groups at comparison schools. The issue is linked with the kind of comparison sites chosen for the study, an issue to be discussed further below. For the moment, it suffices to say that sites already having ongoing programs would not be amenable to having their students randomly allocated to treatment and control conditions, and so already-existing comparison sites and controls cannot practically be combined.

In our judgment, any external validity issues that might be affected by the type of comparison site are outweighed by the sacrifice of internal validity that would result if controls were omitted. It seems inevitable that this sort of

field experiment will be performed on "intact groups"--i.e., that sites or schools rather than individuals will be randomly assigned to various treatments. Without same-school control groups, then, treatment effects are confounded with sampling error differences between schools or sites.

A concern was expressed by the customer that spillover from programs might contaminate controls, and so make the programs seem less effective. Such effects are logically possible, and if they exist, they constitute a serious threat to the validity of the experiment. I know of no empirical support for the contention that these effects exist, but neither is there any assurance that they will not, for programs of the type that were to be developed for PVS. It seems advisable to include controls in schools that do not have intervention programs, as a safeguard. It must be kept in mind for all analyses, however, that the other-school controls, like the nonrandom controls mentioned above, are "intact groups," rather than groups to which each individual is randomly assigned. Such groups have been the center of much controversy in the research literature because they do not satisfy the independence assumption underlying nearly all statistical analyses.

COMPARISON SITES

Just as the need for controls follows from the experimental requirement to rule out rival hypotheses that could account for differences, the need for comparisons follows from the practical criterion that a new intervention should be evaluated against the one it is intended to supplant. The comparison condition is thus a kind of control for what would occur without the innovation, and the kind of comparison site that is required for a study depends upon the way the evaluation questions are formulated

If the question is, "Compared to the existing projects, how well does the new intervention function?" then existing program sites necessarily have to be used. If the question is "In the future, given a choice between starting a project in the old way or implementing the intervention(s) developed in this study, is there a reason to prefer one over the other?" then new sites should be used for the comparison condition as well as for the new intervention condition(s).

The latter form of the question seems more appropriate for policy decisions to be made in the future about the value of a program package. It also makes a better research design possible; if new sites are used, then assignment of sites to conditions (intervention or comparison) can be randomized, and both kinds of sites can have students divided into treatment and control grpups. If comparing performance to existing sites is also of crucial interest, it is possible to use both old and new sites without sacrificing the rigor of the new-site comparisons. If such data are gathered, it is to be expected that comparative analyses will be conducted in which the intervention sites are contrasted with the old comparison sites, but is should be noted that these contrasts are especially vulnerable to "rival hypotheses" about the reasons for differences. The shortcomings of the old comparison sites will be discussed briefly here.

First, if old comparison sites are used, it is necessary to match the (new) intervention sites to them. This requires selecting the old sites first and identifying and selecting similar sites afterward for the intervention. The problem of finding adequately matched intervention sites could be very complex.

Even with careful selection it is clear from the outset that the old comparison sites would differ from the new intervention sites in to ways: whatever selection factor differentiates sites with projects from those without them, and the effects of project experience. This sampling error (i.e., failure to select new sites like the old ones) is one of the major threats to the internal validity of an experiment. Internal validity is further threatened, as mentioned previously, by the lack of same-school control groups, which cannot be employed with old comparison sites. If both old and new comparison sites are used, it is important to recognize that the contrasts involving the old sites are subject to all these threats to internal validity.

LONGITUDINAL ASPECTS

The Planned Variations Study was to be a multi-year project, with a summative phase extending over a two-year period. Time-related questions for such a study require attention to the kind of sample to be drawn for each year and the analytical comparisons to be made.

Timing and analysis of pretests and posttests are longitudinal issues, but have no bearing on the Year 1-Year 2 comparisons. They are, therefore, not discussed here. Timing of tests is primarily a practical matter, requiring coordination with the schools' schedules. The ways of answering pretests and posttests will be treated in the section on analysis.

Assessment of fade-out or residual effects of the programs was called for in the RFP. This poses a practical design problem. It does not seem to be legally or ethically possible to accept students into programs and then, after a period of participation, to exclude them from services, in order to measure fadeout or residual effects. An alternative is to follow up those students who drop out "voluntarily" (i.e., their leaving is not initiated by the program). This approach has the disadvantage of possible bias in the sample--leavers may be systematically different from those who remain--as well as probable small sample size.

Both of these approaches would yield very limited data. If the purpose of studying fadeout is to obtain an optimal solution in the sense that fadeout could be minimized (an assumption on our part--this purpose has not been stated), then a thorough study would examine such variables as entry skill level, amount of exposure, and exit skill level in relation to loss subsequent to termination. Either manipulation or selection on such variables implies a large-scale effort for such a study.

If a strong data base is needed for firm conclusions about fadeout, it should be recognized that such a data base will be very costly and will be obtainable only through a long-term effort. My recommendation would be to concentrate first on creating an effective intervention and establishing its effectiveness through a comparative analysis of a handfull of approaches. In the event one or more of these approaches prove effective, followup studies could be commissioned <u>for just the effective ones</u>, to examine the fadeout question. It seems pointless to study the fadeout of an effect until we know that it exists in the first place.

The study also calls for analysis of data for both years. These analyses can serve to confirm the stability of results over the two years. If results should be different for the two years, then it would be useful to compare second-year scores of students who enter the program in the second year to first-year scores of those who enter in the first year. For this reason, it is desirable to add students at each grade level in the second year. By making all the appropriate comparisons, it would then be possible to assess cumulative intervention effects, grade effects, and historical (cohort) effects. Comparison of second-year scores of, say, tenth-graders who entered the program in the first year with their grademates who entered in the second year would indicate the impact of two years of exposure as compared to one year. Comparing first-year and second-year scores of the students who entered in the first year would provide the standard longitudinal analysis, which confounds program effects, maturation, and any other differences between the general environment in the first and second years. An assessment of these general differences between the first and second years can be made through a comparison of the first-year scores of the students entering in the first year with the second-year scores of those entering in the second year.

ANALYSES OF STUDENT DATA

This section deals with several facets of the most complex part of the analysis effort, i.e, the analysis of student data. Other parts of the analysis, such as analyses of process or context data, are relatively direct readings of the phenomena that are under scrutiny, while the analyses of student data examine information about student performance in order to form inferences about the programs to which the students are exposed. The discussion below is divided into five general categories: definitions, metrics (of program impact), adjustments (of student outcomes), manipulation versus observation, and units of analysis.

DEFINITIONS

Four terms are crucial to understanding the analyses: outcome, benefit, resources, and cost. Each of these is discussed in the following paragraphs.

Outcomes are, in the most general sense, any useful or significant result of a process, such as, in the case at hand, the compensatory programs. Inconveniently, the goals of these programs are so future-oriented that there is no way to determine in the short term whether those goals are being met. It is therefore necessary to deal primarily with surrogates for educational achievement. The four categories of variables that were identified for PVS were basic skills, study skills, affective variables, and behavioral variables.

These categories seem to provide adequate coverage of the spectrum, but identification of variables within each category that predict academic success is a crucial step. It would be worthwhile to develop a predictive system, so that the short-term surrogate variables can be selected on the basis of an empirically-verified relationship to the ultimate outcomes for which they stand.

The concept of benefit differs from that of outcome in that it implies the application of a standard. Students benefit from a program to the degree that they meet criteria for success. Dealing with a success index instead of outcome scores requires that a set of criteria be developed by which scores can be transformed to one or more category variables (1 or 0, depending upon whether the outcome score is above or below the criterion level). The proportion of students

who meet the sucess criteria may yield a different picture of the impact of a program than analyses on outcome means.

Resources, as the term applies in PVS and in similar studies, are program services, materials, and facilities that are brought to bear on students. Because there are a number of these elements allocated in varying amounts to students, a profile of resources is needed to describe a given student's exposure fully. The point here is that the statement that a student had ten hours of exposure to program resources in a given week provides very little information unless it is accompanied by information on the types of resource.

It is difficult to pool profiles in any meaningful way, however, and we need to work with a summary measure of resource exposure. A useful metric is obtained by weighing each resource by its cost, e.g., pro-rata cost of a reading lab for one hour, hourly salary of a teacher aide or a reading remediation specialist, or remedial math materials consumed in an hour. This measure serves an an index of intensity even for a number of components in combination.

Cost is a surprisingly ambiguous term. One version of program cost could be obtained by aggregating the exposure cost data referred to above. But, cost of exposure to a component is computed on an average cost basis, not from actual cost at that site on that day. The costing procedure is therefore not intended to lead to an accurate aggregate, but to provide a measure of intensity of service. Another interpretation of cost is actual program expenditure. This measure is useful in comparing programs that differ in both expenditure level and effectiveness, but cannot be used (given existing accounting procedures) to differentiate program service allocations within a program.

Differences in costs of the same services from one program site to another can render expenditure comparisons misleading, however. Adjusting for cost-of-living (or cost-of-serving) differences between sites provides a cost metric that is more equitable as an index of size of project effort.

Finally, there are cost projections that estimate how much the program would cost to implement under varying conditions, particularly for different client loads. These estimates are useful in determining optimal conditions for implementation, once the program has been tested and disseminated.

Each of these cost indices has a special usefulness, and none can take the place of the others. It is therefore necessary to perform a variety of cost analyses, using several cost indices, to answer all the pertinent cost questions for studies like PVS.

UNITS OF ANALYSIS

In a very important sense, the intervention treatments of PVS would have to be treated as units of analysis. Decisions about the future implementation of the projects should be guided by data indicating the degree of sucess enjoyed by the project during the field test of summative evaluation. For any analysis of data, then, the results must be aggregated, at some stage, to the project level.

The other side of the issue is that projects of this sort are designed to benefit students. Simple aggregations such as mean growth on some measure can be misleading, being based upon the assumption that an increase of a given size has the same utility, or constitutes the same benefit, for all students. On closer examination, this is clearly not the case; students benefit when they can meet criteria for progress through the educational system, whether the increment needed is large or small. For a program that has clear objectives it is possible (and certainly worthwhile) to define criterion levels for the outcome measures. Application of this standard yields a success index by which each individual student can be categorized (e.g., pass/fail). The successes can then be aggregated to the project level as a count or a prercentage.

Although the procedure just described leads to a more meaningful conclusion than analyses performed on outcome means, it involves a step that is open to criticism. The standards that are applied must necessarily include a judgment component--there is no single right answer to the problem of how much students should know in each area. Potential users of the information reported from such studies may

have different standards to apply. We know of no way to develop a benefit metric without using some standard of utility that may run afoul of some readers. For this reason it seems necessary to use both the benefit metric and the outcome metric, which, while less meaningful, is more objective.

INDICATORS OF PROGRAM SUCCESS

The most straightforward indicator of program success is the degree to which it enhances student outcomes: their performance, their feeling about school, their decisions about directions to take following the school experience (whether secondary or postsecondary education), etc. Selection of the particular variables is an important part of the design/analysis process, but cannot be described in detail for the general case of field experiments. The specific data requirements of the study and an assessment of the candidate variables determine what variables should be measured. In contract research the identification of variables necessarily involves both the contractor and the customer. For example, the contractor would typically draw up a list of candidates that are measurable, are relevant to the study issues, and (if possible) have been empirically validated in similar contexts. The tentative list would then be submitted to the customer for approval before proceeding with subsequent tasks such as instrument construction or analysis specifications. The customer's input is crucial at this point to ensure that the study's policy issues are addressed.

When the set of variables has been defined and measurement procedures have been worked out, it is necessary to decide what analyses will be applied to the resulting data. The armamentarium of analyses available for this sort of research problem places heavy emphasis on group means, whether they are mean post-treatment scores or means on some sort of change score. The main problem with this analytic approach is tht it rests on the not-altogether-tenable assumption that a unit of gain is equally important regardless of where it occurs on the scale. This seems unsatisfactory; it is intuitively incorrect to assign, say, a ten-point gain the same importance whether it brings the student above the criteria for acceptance or not.

One alternative, suggested by Melvin Novick,* is to transform the scores to a utility scale where a utility is, by definition, equally valuable regardless of level. Unfortunately, the procedure recommended for generating the utility scale requires a separate side study whenever such a transformation scale is to be developed, and so may be too cumbersome and costly for some research endeavors.

A less complicated approach is to set criterion levels for student success (not an easy task, but one that must eventually be done in any event) and use a count (or relative frequency) of those who meet the criteria as an indicator of project success. This step serves to convert outcomes into a form of benefits indicator, as previously stated, by applying a standard. It is, consequently, fraught with subjectivity. The values upon which this subjective judgment is based should, it appears, be those of the customer or his chosen referee (such as an advisory group), to ensure that the standards accurately represent program goals. Besides its subjectivity, this approach also has a disadvantage relative to the utility-theory approach in that it yields only categorical data, but such data may be quite sufficient for the evaluation purpose being discussed.

For both outcome and benefit, it is important to note that it is not simply the absolute level that is important, but the enhancement of outcome or benefit, relative to alternative approaches. The design of studies like PVS must provide for this relative emphasis. There appears to be no adequate substitute for a genuine field experiment with appropriate controls. In the case of PVS, the design included randomized controls who would receive no intervention services and randomly-assigned comparison schools where programs would be operated under existing Special Services regulations.

Enhancement of the ratio between outcome and cost or between benefit and cost are important additional indicators of program success. Although the importance of these indicators is obvious, the hurdles may not be. The simplest kind of assessment is to look at costs for these programs that are shown to be effective.

---

*Oral communication, 1976.

This can be carried out with little extra effort. If critical mass and the functional relationship between cost and effectiveness are desired, however, cost would have to be manipulated as an independent variable or observed in its natural variations. The latter approach would be unsatisfactory because the range that is of potential interest would not necessarily occur naturally. In fact, other funding factors would tend to make program costs fairly uniform, and the natural occurrence of the range of cost that should be studied would be very unlikely to just happen. Within these limitations of the range, cost/effectiveness can be examined without overly complicating the study design. The various interpretations of what is meant by cost were discussed previously in this document.

ADJUSTMENTS OR CORRECTIONS TO SCORES

A field-experimental study like PVS involves the measurement of complex differential quantities. The requirement that the impact of a program be measured implies longitudinal assessment of change over time, and the need for assessing this change relative to changes attributable to alternative programs implies examining (conceptually, at least) differences between changes attributable to the programs. This difference between differences leads to a category of analysis that has been the subject of a great deal of controversy and the cause of a great deal of inconsistency between precept and practice. Some alternative approaches to the problem will be discussed briefly here.

UNCORRECTED POST-TREATMENT SCORES

Well-known papers by Lord (1967)[1] and by Cronbach and Furby (1970)[2] discussed a number of "correction" procedures, all of which were intended to provide an approximation to the ideal "all other things being equal." Some of these will be discussed below. The point to be made here is that these papers concluded by

---

[1] Lord, F. M. A Paradox in The Interpretation of Group Comparisons. Psychological Bulletin, 1967, 68, 304-305.

[2] Cronbach, L. J. and Furby, L. How Should We Measure "Change"--Or Should We? Psychological Bulletin, 1970, 74, 68-80.

recommending experimental rather than statistical control of the "other things." The recommendation was that randomly-assigned controls be used (whenever possible), and that post-treatment scores be compared. The logic seems to be accepted almost universally, but the suggested approach is rarely used in educational research. In part, this can be attributed to the fact that few studies have the option of including legitimate randomized controls, but it is also possible, in informal conversations with researchers, to detect an attitude of "Well, yes, the math and logic make sense, but nobody (i.e., of the contract monitors) would accept a study that made no use of the pretest data."

Fortunately, it is not necessary to choose between experimental and statistical control. In fact, Lord's point was that statistical "corrections" such as analysis of covariance could not make up for the absence of control groups. A similar point has been clearly stated by Campbell and his associates (e.g., Campbell and Stanley, 1963)*, who always give first priority to true experimental designs, but the point is often lost in the mass of discussion of substitute "quasi-experimental" designs, which is the "next best" solution. It is possible, then, to use statistical correction procedures when controls have been randomly selected, and such corrections are more appropriately used in this way than as a substitute for randomization. The selection of a correction procedure should be informed by a careful analysis of the reason for correcting. Some of the issues to be considered follow.

INDICES OF GROWTH USING PRE-TREATMENT MEASURES

Two major categories of issues are concerned with (1) what the scores are being corrected to (or for) and (2) how the resulting scores correlate with other things. In my opinion, the former set of issues are the most important ones, because they are directly related to the meaning of the scores obtained. The intercorrelations between corrected posttest scores and other scores, particularly

---

*Campbell, D. T. and Stanley, J. C. Experimental and Quasi-Experimental Designs for Research on Teaching. In N. L. Gage (Ed.) Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

pretest scores, are important primarily in that they may be based partly upon irrelevant relationships, such as those based upon irrelevant features that are shared by the pretest and posttest. Other issues related to intercorrelations have to do with the reliability of the derived score and the derived score's accuracy in terms of being representative of a true change score. Cronbach and Furby (1970, op. cit.) have analyzed the uses to which gain scores had been put, and suggested one or more other approaches for each kind of question.

One of the uses they discussed fits the PVS study; they suggested using true score estimates for the comparison of nonrandomly formed treatment groups. (The same argument seems to apply equally where randomization procedures have been applied but a comparison of the two groups indicates that the randomization has failed to equalize them.) In any case, the validity of corrections is not destroyed if randomization is satisfactory--they are simply not profitable.

For studies like PVS, it is not only necessary to know which treatment leads to superior outcomes, it is also important to know how much improvement can be expected for individual students, given their initial description (including, but not limited to, pretest scores). This projection calls for a highly reliable variable that represents the relationship between initial-state measures and posttest measures. If a linear regression captures most of the variance in the criterion measures (or a composite of them), then the regression slope of each treatment would be useful. First, it would be directly readable as a group's growth index. Second, it would yield a prediction of the outcome criterion variable for any initial-state score vector.

The major disadvantage inherent in this approach is that a regression slope doesn't look much like a change score and probably would not be generally accepted as the sole answer to the question: "How well do the programs work, for what kinds of students?" Another approach that was to be developed for PVS combines some features of Cronbach and Furby's (1970, op. cit.) true score

estimation procedure and Campbell and Erlebacher (1970)[1] regression projection or experimental regression design, presented in some detail by Sween (1971)[2]. Like the true score estimate, it utilizes a number of initial-state variables to generate a linear prediction of the criterion variable. Like the Campbell-Erlebacher approach, it generates predictions using the control group's pretest data to estimate what "would have occurred" if the intervention had been omitted.

The approach is to develop the regression expression relating the control group's initial-state-variable scores to their criterion scores, apply this regression to the treatment group's initial-state-variable scores to obtain an estimate of their scores without the intervention, and obtain residualized scores by subtracting the estimates from the actual scores. These residualized scores would have a clear meaning: the difference between the actual outcome with the intervention and the best available estimate of the outcome that would have been obtained without it. The predictor variables would be screened beforehand (e.g., during the formative year) to find those that best account for the variance in the criterion variable. Although Campbell and Erlebacher's model was designed to deal with non-equivalent groups, the rationale generalizes to equivalent groups. The only question is whether it "buys" anything for the extra effort. This can best be answered by examining the pre-treatment (initial-state) scores of both treatment and control groups. If they are satisfactorily equivalent, post-treatment scores alone suffice; otherwise, the residualizing techniques can be used if the assumption that the regression slopes for the treatment and control groups would have been equal without the intervention is tenable.

MANIPULATION VERSUS OBSERVATION

A study that calls for the creation of an effective intervention, followed by a test of that intervention in sites (as required by PVS) must balance two concerns

---

[1] Campbell, D. T. and Erlebacher, A. How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful. In J. Hellmuth (Ed.) Compensatory Education: A National Debate, Vol. 3 Disadvantaged Child. New York: Brunner/Mazel, 1970.

[2] Sween, J. A., The Experimental Regression Design: An Inquiry into the Feasibility of Nonrandom Treatment Allocation. Ph.D. Dissertation, 1971.

that are potentially in conflict: the need for an orderly study design that provides a means for measuring the effectiveness of the intervention and the need of the intervention designers to optimize the potential of the intervention. We believe priority should be given to the development of the intervention; it seems pointless to design a study if the object of that study (i.e., the intervention) is disrupted by the design. In such a case it would be impossible to determine whether the originally intended intervention would have been successful or not. The general ground rule we observe, then, is to manipulate, control, block, etc., wherever possible without interfering with the process under study.

The clearest example of this conflict between study design and intervention design and development was in the issue of randomly assigned treatment variations versus naturally occurring treatment variations. In the former, some levels of the various treatment components would have been chosen, presumably on some rational basis, and the students participating at each project site would have been subdivided randomly and assigned to the treatment variations. This approach would lead to serious logistical problems as well as sample size difficulties caused by subdividing groups, but an even more important disadvantage to the approach is that the assignment of a treatment sequence to any given student is determined by random selection (dictated by the study design) rather than on the basis of the project's diagnosis of need, as preferred by the program developer.

In such a conflict, it is necessary to balance how much the programs would be weakened if the diagnostic-prescriptive function were abandoned, against how much the evaluation would be weakened if the treatment sequences were not manipulated. In PVS it was decided that assignment to component activities according to the results of a need assessment was an essential part of the development strategy, and could not be excluded without deviating from the instructional philosophy around which the programs were being designed. This decision presented to the study design staff the challenge of finding a way to compare effectiveness of a variety of individually prescribed treatment sequences to (1) that of a comparison treatment (using standard Special Programs approaches) and (2) that of a school program with no special compensatory activities.

The simplest approach to the problem is to ignore the variations in individual student experience within a program, and simply compare post-treatment outcome means for the intervention groups with the post-treatment outcome means for the non-intervention groups. This approach is defensible as a way to evaluate the entire program activity package, including the assessment-prescription component, disregarding any internal heterogeneity.

The above approach has two shortcomings: (1) it leaves what may be a great deal of internal heterogeneity to be absorbed as error variance, thus lowering the sensitivity of the study, and (2) it treats the program as an unseparable unit, providing no basis for evaluating the impact of a component. A second approach overcomes both of these problems. Projections are obtained for scores on the outcome variables for each treatment-group student, using the regression model from the control group and the initial-state predictor scores of the treatment-group student. These projections represent a best estimate of "what would have occurred" without the intervention treatment experienced by that student. The score obtained by subtracting the projected score from the score actually obtained can then be regressed on the activity exposure variables to obtain a relational analysis describing the contribution of each activity to each outcome. This approach does not yield the degree of confidence that results from experimental manipulation of the variables, but is the kind of exploratory analysis that is well suited to the state of knowledge in this field. A description of the relationships that appear in such natural variations could form the basis for a later experimental confirmation. The experimental-manipulative approach would be premature at this time, however; there is not an adequate basis (either empirical or theoretical) for selecting appropriate levels for the manipulable variables.

OTHER CATEGORIES OF EVALUATION DATA

Unlike products in many other fields, educational programs cannot be evaluated as objects that have certain desirable characteristics. The prevalent view of them is perhaps most appropriately represented as an operator whose operand is, in most cases, the student. In other terms, the student is somewhere in a state space and the purpose of the program is to induce a transition to some other part of the state space. It follows, then, that the greater part of the evaluation effort for such programs is devoted to analysis of changes in students. (As was pointed out in a previous section, the study of change is a complex issue and need not always involve the observation of change scores.)

The problem is even less straightforward than the above paragraph would suggest, however. Since the program is the product that can be modified or improved on the basis of field test data, the causal links between identifiable procedures/activities/experiences and outcomes need to be identified as well as possible. Part of this task was covered in the section on student outcome analyses because the analyses specifically included student outcomes. Other checks are also needed, however, to ensure adequate representation of the relevant environmental factors that constrain the changes the student undergoes. Examination of the details of the process to which the student is exposed is essential--in fact, much of the process analysis is dealt with in the relational analysis of student outcomes--and observational data are of great value in the identification of any unanticipated characteristics of the process. A part of the process analysis that probably needs to be treated as a separate category is the assessment of fidelity of implementation. The objective here is simply to ensure that the program as implemented corresponds closely to the program as conceptualized. Similarly, contextual characteristics need to be identified, and may have to be considered in an ad hoc manner to account for perturbations in an otherwise orderly data pattern. Finally, the desire to capitalize on any serendipitous outcomes of programs leads to the examination of other programs and other characteristics of the program environment that change concurrently with (and, therefore, seemingly because of) the introduction of the program. These are sometimes treated in the "spinoff" category. Problems specific to each of these categories will be treated below.

PROCESS

The study of process involves two major emphases: an accounting function, in which kinds and amounts (of each kind) of service are recorded for each individual student; and an auditing function, in which the actual services and ways of allocating them are checked against specifications to determine fidelity of implementation. The latter will be discussed separately, below.

The accounting function provides data that fit into the student outcome analyses, discussed previously. The specification of process data to be collected and methods for data collection were not covered there, however, and constitute a special kind of problem, to be discussed here. In general terms, the kinds of information needed can be defined at the outset: kinds of services or activities experienced by some or all of the students, as determined by the intervention program; and amount (time) of exposure to each. While this step is simple and immediate, the specification of the variables must be deferred until program structure is sufficiently defined in the course of program development. Once the components are defined and a schedule of delivery is specified, it is possible to specify the variables and, tentatively, the level of precision with which each should (or can feasibly) be measured.

Similarly, it is possible to define at the outset the set of measurement/data recording techniques that can be used, but decisions as to which technique(s) will be used for which variables must wait until the variables and their requisite precision are determined. The decision about techniques must, of course, be shared by those responsible for program implementation.

ASSESSMENT OF IMPLEMENTATION

Monitoring of the fidelity of implementation of a program is also based upon process data, but the objectives of this activity are different from those of the process analyses already discussed. The monitoring function provides, primarily, feedback to the implementer that can serve as a basis for correcting the implementation. If it should prove impossible to implement the program with sufficient fidelity at a given site, the monitoring information could be used as a justification for discounting the data from that site, but barring such an

extreme, the monitoring data is entirely for formative (real-time corrective) rather than summative purposes.

CONTEXT

The assessment of program context should identify context characteristics that interact with programs to affect the students' growth. This information can be used to moderate conclusions about differences among sites if context variations seem to be associated with biases in the results. This is the least satisfying purpose for context analysis because it typically has to be done on an ad hoc basis, using some irregularity observed in the context to account for differences between the site with the irrgularity and another site. Such explanations do not add to the store of scientific knowledge, but may nonetheless be useful in keeping bad data out of that accumulation of information.

A more structured use of context information is as a basis for selecting sites where programs will be installed after their evaluation. Within the limits of practicality, the best way to establish the potency of a context characteristic (in interaction with the program) is to include it as a block in a block x treatments design. There is clearly a cost-benefit tradeoff here, however, since blocking is both the cleanest and the most expensive approach. Each blocking factor in the design multiplies the number of sites required by the number of levels on that factor or, alternatively, introduces within-site blocking. The latter is frequently not logistically feasible, introduces potential cross-contamination, and reduces the available sample size per condition. Finally, optimal blocking requires prior knowledge of the characteristics that are most likely to be potent variables.

Alternatively, measures can be taken on a large number of characteristics, and the naturally occurring variations in these variables from one site to the next can be used as covariates (or predictions). A much larger number of potentially potent predictors can be investigated in this way, at relatively low cost. If confirmation of these variables' predictive power is needed, a small number could be selected and confirmed in a blocking design. The point is that the discovery of the important context variables is best handled through regression/correlation

procedures, and for newly created programs it is discovery, rather than confirmation, that has to be accomplished.

The selection of characteristics to be investigated should be guided by the feasibility of measuring them within the constraints on the study being designed. Perhaps even more importantly, the characteristics should have policy relevance--some potential must exist for useful action to be taken on the basis of the information. If there is a way to (1) select sites on the basis of the study's findings, (2) modify the programs to confirm to site characteristics, or (3) change the sites to eliminate the characteristics that interact negatively with the programs, then it is worthwhile to know about those site characteristics. Any characteristics that cannot be controlled or corrected for in some way are not justifiable for inclusion in a context analysis as a part of an evaluation study. (This conclusion presupposes that the purpose of an evaluation is to provide information to guide decisions).

SPINOFFS/SPILLOVER

It is sometimes of interest to determine how and to what extent programs affect other activities. This is particularly important if there is reason to expect harmful side effects that would outweigh the salutary outcomes of the program, or if a significant amount of the program's value is in its effect upon the general population of students (or other target groups) rather than just those who are designated recipients of its services. The search for such effects is difficult to structure unless there are specific areas where spinoff or spillover effects are anticipated. These expectations typically arise from experience with similar programs or a logical extrapolation from the characteristics of the program. In either case, the foundations for such expectations is better after program development has been completed. In the absence of a well-structured approach, it is possible to identify, at least, areas of potential impact and informants, and to gather information through unstructured interviews.

SUMMARY AND CONCLUSION

The preceding sections have discussed a sequence of activities by which design and analysis plans can be generated for a study like PVS. A four-step procedure for specifying design and analysis approaches was described, and some special concerns in the design and analysis plans were presented. The discussion had, for the most part, a "how to" or "how not to" flavor, and there was some emphasis upon problems encountered and suggestions for approaches that would avoid or alleviate them. In most cases the suggestions are directed to project planners whose duty it is to interact with the customer, but they occasionally address action that can only be initiated by the customer.

Because the report deals with experiences relevant to design and analysis for PVS-like projects, only the planning steps were discussed. Subsequent steps, such as generating detailed Analysis Specifications and an Analysis System, would be carried out after the planning period, and so are not a part of the experience base from which this report was written.

A great deal more could be written about the problems involved in initiating a jointly-funded endeavor like PVS, including the planning of the cooperative effort prior to writing an RFP and the subsequent give-and-take of ideas across disciplinary as well as departmental boundaries. While these are no doubt crucial and more intriguing facets of such efforts, that chapter would have to be written by those who experienced it. The problems that confront the researcher are the few that "get through the net." Those problems that have seemed most important to planning the design and analysis of PVS were the difficulty in getting acceptance on the steps in the planning phase and the mistaken assumptions about the availability of adequate data for the Target Population Analysis. These and various technical problems and their solutions were discussed.

APPENDIX

STUDY QUESTIONS

STUDY QUESTIONS

1.0 STUDENT OUTCOME QUESTIONS

1.1 Effects of Treatments Relative to Controls

What are the effects of each treatment on student behavioral, cognitive, and affective outcones, relative to the regular school program, with different durations of exposure, as assessed through a randomized controlled experiment?

1.1.1 Same-School Controls

1.1.1.1 Mean Outcome Scores

1.1.1.1.1 How do mean outcome scores for intervention and comparison treatment students compare to those for control students at the same schools after one year?

1.1.1.1.2 How do mean outcome scores for intervention and comparison treatment students compare to those for control students at the same schools after two years?

1.1.1.1.3 How does the impact of each treatment on outcome means relative to same-school controls change from the first year of participation to the second year?

1.1.1.2 Adjusted Outcome Scores

1.1.1.2.1 What are the adjusted mean outcome scores for intervention and comparison treatment students after one year of student participation?

1.1.1.2.2 What are the adjusted mean outcome scores for intervention and comparison treatment students after two years of student participation?

1.1.1.2.3 How do the adjusted mean outcome scores change from the first year of participation to the second?

1.1.1.2.4 How are the student characteristics (which are used to adjust outcome scores in the previous analyses) related to the outcome scores?

1.1.1.3 Success Rates

1.1.1.3.1 How do the treatment and same-school control groups differ in success rate after one year of student participation?

1.1.1.3.2 How do the treatment and same-school control groups differ in success rate after two years of student participation?

1.1.1.3.3 How does the impact of each treatment on success rate (relative to same-school controls) change from the first year of participation to the second?

1.1.2 Other-School Controls

1.1.2.1 Mean Outcome Scores

1.1.2.1.1 How do mean outcome scores for intervention and comparison treatment students compare to those for control students at different but comparable schools after one year of student participation?

1.1.2.1.2 How do mean outcome scores for intervention and comparison treatment students compare to those for control students at different but comparable schools after two years of student participation?

1.1.2.1.3 How does the impact of each treatment on outcome means, relative to different-school controls, change from the student's first year of participation to the second?

1.1.2.2 Success Rates

1.1.2.2.1 How do the treatment and different-school control groups differ in success rate after one year of student participation?

1.1.2.2.2 How do the treatment and different-school control groups differ in success rate after two years of student participation?

1.1.2.2.3 How does the impact of each treatment on success rate (relative to different-school controls) change from the first year of participation to the second?

1.1.3 Comparisons Between Controls

1.1.3.1 Mean Outcome Scores

1.1.3.1.1 How do mean outcome scores for same-school control students (the spill-over group) compare to those for different-school control students after one year of student participation?

1.1.3.1.2 How do mean outcome scores for same-school control students (the spill-over group) compare to those for different-school control students after two years of student participation?

1.1.3.1.3 How does the difference between mean outcome scores for the two kinds of control students change from the first year to the second?

1.1.3.2 Success Rates

1.1.3.2.1 How do the same-school controls and the different-school controls differ in success rate after one year of student participation?

1.1.3.2.2 How do the same-school controls and the different-school controls differ in success rate after two years of student participation?

1.1.3.2.3 How do the differences between the success rates of the same-school controls and different-school controls change from the first program year to the second?

1.1.3.3 Regression Formulas

1.1.3.3.1 How do the regression formulas derived (for the adjustment of treatment scores) from the same-school and different-school control groups differ? Are they practically equivalent, or should each be used in the analyses?

1.2 Effects of Treatments Relative to Each Other

How do the effects of the treatments on student behavioral, cognitive, and affective outcomes compare with each other after one year and after two years of implementation?

1.2.1 Mean Outcome Scores

1.2.1.1 How do mean outcome scores for the interventions at each level compare with each other after one year of student participation?

1.2.1.2 Ho do mean outcome scores for the interventions at each level compare with each other after two years of student participation?

1.2.1.3 How does the difference betwen (or among) mean outcome scores for the interventions change from the student's first year of participation to the second?

1.2.2 Adjusted Outcome Scores

1.2.2.1 How do the adjusted mean outcome scores for the interventions at each level compare with each other after one year of student participation?

1.2.2.2 How do the adjusted mean outcome scores for the interventions at each level compare with each other after two years of student participation?

1.2.2.3 How does the difference between (or amont) adjusted mean outcome scores for the interventions change from the student's first year of participation to the second?

1.2.3 Success Rates

1.2.3.1 How do the success rates for the interventions at each level compare with each other after one year of student participation?

1.2.3.2 How do the success rates for the interventions at each level compare with each other after two years of student participation?

1.2.3.3 How does the difference between (or among) success rates for the interventions change from the student's first year of participation to the second?

1.3 Effects of Interventions Relative to Special Services Programs

How do the effects of each intervention on student behavioral, cognitive, and affective/outcomes compare with the effects of newly initiated Special Programs on comparable students after one year and after two years of implementation?

1.3.1 Mean Outcome Scores

1.3.1.1 How do the effects of each intervention condition on mean outcome scores compare with the effects of the comparison programs after one year of student participation?

1.3.1.2 How do the effects of each intervention condition on mean outcome scores compare with the effects of the comparison programs after two years of student participation?

1.3.1.3 How do the differences between means for each intervention condition and the comparison programs change from the first year of student participation to the second?

1.3.2 Adjusted Outcome Scores

1.3.2.1 How do the effects of each intervention condition on adjusted outcome scores compare with the effects of the comparison programs after one year of student participation?

1.3.2.2 How do the effects of each intervention condition on adjusted outcome scores compare with the effects of the comparison programs after two years of the student participation?

1.3.2.3 How do the differences between adjusted means for each intervention condition and the comparison programs change from the first year of student participation to the second?

1.3.3 Success Rates

1.3.3.1 How do the effects of each intervention condition on success rates compare with the effects of the comparison programs after one year of student participation?

1.3.3.2 How do the effects of each intervention condition on success rates compare with the effects of the comparison programs after two years of student participation?

1.3.3.3 How do the differences between success rates for each intervention condition condition and the comparison programs change from the first year of student participation to the second?

1.4 Residual Effects of Interventions

Within the time span of the Planned Variations Study, what are the residual intended or unintended effects, or what is the "fade-out" of effects, on student outcomes, after treatment termination, and what experiences are the determinants of these effects? (Provision will also have to be made in the planning of the study to continue follow-up after completion of this study.)

1.4.1 How do second-year follow-up outcome scores on leavers from the intervention programs differ from their predicted scores based on continuing program participant scores at the same time? (Fade-out)

1.4.2 How do second-year follow-up outcome scores on leavers from the intervention programs differ from their predicted scores based on control-student regressions? (Residual)

1.4.3 How are program exposure variables related to fade-out?

1.4.4 How are program exposure variables related to residual effects?

1.4.5 How do leavers compare with intervention participants with respect to such behavioral outcomes as college entry or college completion at various intervals after the second summative year?

1.4.6 How do leavers compare with controls with respect to such behavioral outcomes as college entry or college completion at various intervals after the second summative year?

1.5 Leaver Characteristics

What characteristics differentiate the intervention and comparison program dropouts from students who remain in the programs? Can reasons for leaving be identified?

1.5.1 Which of the student background, pretest, and needs assessment variables differentiate leavers from students who remain in the intervention programs?

1.5.2 What reasons do students give for leaving the intervention programs?

2.0 DESCRIPTION OF PROCESS AND CONTENT

2.1 Receipt of Program Services in Interventions

What combinations of program services are actually delivered to students participating in the intervention programs?

2.2 Receipt of Program Services in Special Programs

What combinations of program services are actually delivered to students participating in the UB and SSDS programs?

2.3 Comparison of PVS Interventions With Special Programs

What characteristics differentiate the processes and contents of the Planned Variations Study interventions from those of the comparison (Upward Bound and Special Services for Disadvantaged Students) programs?

3.0 ANALYSES OF PROGRAM COMPONENT EFFECTS

How does exposure to, or participation in, program activities influence student performance?

3.1 Influence of Components on Outcome Scores

How are the scores on the outcome measures of PVS influenced by the amount of exposure of the student to program activities?

3.2 Influence of Components on Likelihood of Student Success

How is probability of success (defined as meeting PVS criteria) influenced by the amount of exposure of the student to program activities?

4.0 COST-OUTCOME ANALYSES

4.1 Actual Program Costs Versus Outcomes

4.1.1 How are actual total program costs related to adjusted mean outcomes?

4.1.2 How are actual per-pupil program costs related to adjusted mean outcomes?

4.2 Maximum-Efficiency Cost Projections Versus Outcomes

4.2.1 How are maximum-efficiency projections of actual total program costs related to adjusted mean outcomes?

4.2.2 How are maximum-efficiency projections of actual-dollar per-pupil program costs related to adjusted mean outcomes?

4.3 Maximum-Efficiency Standardized Cost Projections Versus Outcomes

4.3.1 How are maximum-efficiency projections of standardized total program costs related to adjusted mean outcomes?

4.3.2 How are maximum-efficiency projections of standardized per-pupil program costs related to adjusted mean outcomes?

4.4 Estimated Component Costs Versus Individual Outcome Scores

4.4.1 How are estimated dollar costs of program activity exposure related to individual student success indices?

5.0 COST-BENEFIT ANALYSES

5.1 Actual Program Costs Versus Benefits

5.1.1 How are actual total program costs related to student success rate?

5.1.2 How are actual per-pupil program costs related to student success rate?

5.2 Maximum-Efficiency Projections of Actual Costs Versus Student Success Rates

5.2.1 How are maximum-efficiency projections of actual total program costs related to student success rate?

5.2.2 How are maximum-efficiency projections of actual per-pupil program costs related to student success rate?

5.3 Maximum-Efficiency Projects of Standardized Costs Versus Success Rates

5.3.1 How are maximum-efficiency projections of standardized total program costs related to student success rates?

5.3.2 How are maximum-efficiency projections of standardized per-pupil program costs related to student success rates?

5.4 Estimated Component Costs Versus Individual Student Success

5.4.1 How are estimated dollar costs of program activity exposure related to individual student success indices?

6.0 JUDGED EFFECTIVENESS

6.1 How effective are the PVS interventions in meeting their objectives, in the opinions of project personnel?

6.2 How effective are the PVS interventions in meeting their objectives, in the opinions of institutional personnel?

6.3 How effective are the PVS interventions in meeting their objectives, in the opinions of national personnel?

7.0 CONTEXT EVALUATION

7.1 Cost/Benefits Influence and Relationships

How is the relationship between program costs and benefits influenced by other supplemental compensatory expenditures and by regular school or college expenditures?

7.2 Context Other Than Funding

7.2.1 How does the community context in which projects exist influence their characteristics?

7.2.2 How does the community context in which projects exist influence their effectiveness?

7.2.3 How does the institutional context in which projects exist influence their characteristics?

7.2.4 How does the institutional context in which projects exist influence their effectiveness?

8.0 ASSESSMENT OF PROGRAM IMPLEMENTATION

8.1 Fidelity of Implementation

During the Planned Variations Study, in what ways do the implementations of the interventions at the various sites deviate from plan?

8.2 Implementation Problems

During the Planned Variations Study, what problems are encountered by project staff in implementing the study's interventions?

8.3 Evaluation of Technical and Administrative Support

What technical and administrative support provided to the implementation sites during the study prove to be significant and could subsequently be provided by USOE after the study's conclusion?

8.4 Problems Encountered by SDC

What problems are experienced by the developer (SDC) in implementing the models?

9.0 EFFECTS ON OTHER PROGRAMS

What, if any, impacts do the PVS interventions have on the other programs in the experimental sites?

10.0 CONCLUSIONS AND RECOMMENDATIONS

What interventions work for what students in what contexts?